

# Contents

List of Figures . . . . .	iii
List of Tables . . . . .	iv
<b>1 Introduction</b>	<b>1</b>
1.1 The DANCe project . . . . .	1
<b>2 The Pleiades as a benchmark</b>	<b>3</b>
2.1 Generalities . . . . .	3
2.2 Spatial Distribution . . . . .	4
2.3 Velocity Distribution . . . . .	4
2.4 Luminosity Distribution . . . . .	4
2.5 Mass Distribution . . . . .	4
2.6 The current dynamical scenario . . . . .	4
2.7 The Pleiades DANCe DR2. . . . .	4
<b>3 Bayesian formalism</b>	<b>7</b>
3.1 Introduction to probability theory. . . . .	7
3.2 Bayesian hierarchical Models . . . . .	16
3.3 Modelling the data . . . . .	19
3.4 Priors . . . . .	29
3.5 Sampling the posterior distribution . . . . .	30
3.6 Codes . . . . .	33
<b>4 Results</b>	<b>34</b>
4.1 Performance of the classifier . . . . .	34
4.2 Velocity distribution . . . . .	34

	ii
4.3 Spatial distribution . . . . .	34
4.4 Luminosity distribution . . . . .	34
4.5 Mass distribution . . . . .	34
4.6 The mass distribution on time . . . . .	34
4.7 The phase space . . . . .	34
4.8 Updating the previous knowledge . . . . .	34
<b>5 Conclusions and Future Work</b>	<b>35</b>
<b>A Appendix Chapter Title</b>	<b>36</b>

# List of Figures

2.1	$K$ vs $Y - J$ CMD for the Pleiades candidate members of Bouy et al. (2015b)	5
3.1	PGM representing the parametric inference of a normal distribution.	18
3.2	Normalised mean (left panel) and variance (right panel) of each parameter in our model, as functions of iterations. The normalisation values are the mean and variance of the ensemble of particles positions at the last iteration. Red lines show one and two sigma levels of these normalisation values.	32

# List of Tables

2.1 Pleiades DANCe DR2 properties . . . . . 4

# Chapter 1

## Introduction

Description of:

- Motivation to investigate nearby young open clusters.
- The importance of the IMF
- Current state of knowledge in dynamical simulations.
- The problematic of constraints in dynamical theories.

It must be clear which is the objective

Description of the Nearby open clusters and their properties.

### 1.1 The DANCe project

- List of open clusters in the DANCe project.
- the importance of the pleiades, why we restrict to it.

It must be clear what are the limitations, the boundaries in which the objective will be searched

Description of the current methodologies used to address the question mentioned previously.

- The works of Sarro, Krone-Martins, Malo, Gagne etc.
- The advantages and caveats of the previous methodologies.
- It must be clear the necessity of a new perspective

The proposal we made. The use of Bayesian Hierarchical Models. Benefits and issues of BHM.

Description of the advantages of BHM.

-¿ It must be clear that BHM are the best choice.

Description of the practical issues needed to be solved in order to use BHM.

MCMC techniques and PSO.

-¿ It must be clear that MCMC methods are the best option.

Brief descriptions of our results and how they impact our current knowledge.

-¿ It must be clear that we attained the objective: The pleiades velocity, spatial and mass distributions.

## Chapter 2

# The Pleiades as a benchmark

### 2.1 Generalities

Description of the current knowledge on the pleiades cluster Age, distance, metallicity reddening, number of members, proper motion. Describe what are they.

### 2.1.1 The distance controversy

## 2.2 Spatial Distribution

## 2.3 Velocity Distribution

### 2.3.1 Radial Velocity

## 2.4 Luminosity Distribution

## 2.5 Mass Distribution

## 2.6 The current dynamical scenario

## 2.7 The Pleiades DANCe DR2.

This section must contain a detailed description of the DR2 data.

Table 2.1: Pleiades DANCe DR2 properties

### 2.7.1 Particularities of the Pleiades DANCe DR2

As described in Section 2.7 the Pleiades DANCe DR2 contains astrometric (stellar positions and proper motions) and photometric ( $ugrizYJHK_s$ ) measurements for 1,972,245 objects.

### 2.7.2 Selection of variables

Sarro et al. (2014) demonstrated that the most effective variables for the discrimination of members are the proper motions and the  $riYJHK_s$  bands. However, they excluded the  $r$  band due to its large number of missing values in their training



set. The selection of variables in this work aims at comparing its results with those found by Bouy et al. (2015b) using the  $\mu_\alpha, \mu_\delta, J, H, K_s, i - K_s, Y - J$  variables.

The set of variables used in this work are the stellar positions, the proper motions in right ascension and declination,  $\mu_\alpha, \mu_\delta$ , and the photometric colours and magnitudes,  $i - K_s, Y, J, H, K_s$ . However, in order to compare the results with those of Bouy et al. (2015b), the analysis of the spatial distribution of the Pleiades stellar positions is done independently, in ?.

As described in ?, the photometry is modelled by parametric series of cubic spline. The parameter of this series is the colour index  $i - K_s$  (in the following *CI*). This colour allows the most one-to-one variate-covariate relation. Figure 2.1 shows the colour-magnitude diagram (CMD)  $K$  vs  $Y - J$  the second most one-to-one relation.

Figure 2.1:  $K$  vs  $Y - J$  CMD for the Pleiades candidate members of Bouy et al. (2015b)

### 2.7.3 Data preprocessing

Since both photometry and proper motions carry crucial information for the disentanglement of the cluster population, we restrict the data set to objects with proper motions and at least two observed values in any of our four CMDs:  $Y, J, H, K_s$  vs *CI*. This restriction excludes 22 candidate members of Bouy et al. (2015b), which have only one observed value in the photometry. Furthermore, we restrict the lower limit ( $CI = 0.8$ ) of the colour index to the value of the brightest cluster member. We do not expect to find new bluer members in the bright part of the CMDs. We set the upper limit ( $CI = 8$ ) of the colour index at one magnitude above the colour index of the reddest known cluster member, thus allowing for new discoveries. Due to the sensitivity limits of the DR2 survey in  $i$  and  $K_s$  bands, objects with  $CI > 8$  have  $K_s$  magnitudes  $\geq 16$  mag. These objects are incompatible with the cluster sequence and therefore we discard them a priori as cluster members.

Our current computational constraints and the costly computations associated

to our methodology (described throughout this Sect.), prevent its application to the entire data set. However, since the precision of our methodology, as that of any statistical analysis, increases with the number of independent observations, we find that a size of  $10^5$  source for our data is a reasonable compromise. Although a smaller data set produces faster results, it also renders a less precise model of the field (in the area around the cluster) and therefore, a more contaminated model of the cluster. For these reasons, we restrict our data set to the  $10^5$  objects with highest membership probabilities according to Bouy et al. (2015b). Of this resulting data set, the majority ( $\approx 98\%$ ) are field objects with cluster membership probabilities around zero. Thus, the probability of leaving out a cluster member is negligible. For the remaining of the objects in the Pleiades DANCe DR2, we assign membership probabilities *a posteriori*, once the cluster model is constructed.

## Chapter 3

# Bayesian formalism

This chapter provides a general introduction to probability theory and its application to parametric inference. Since the objective of this work is to infer the probability distributions of the cluster properties (e.g. luminosity and velocity), I give reason to prove that the Bayes' theorem provides the proper framework for the inference of the parameters governing these distributions. Later in this chapter, I describe the reason for which the Bayesian Hierarchical Models are the best option to parametric inference under the Bayesian framework.

In the following Sections I will describe in detail the assumptions I made to model the data and to select the prior distributions. The two final Sections of this Chapter focus on: the practical issues related to the sampling of the cluster distributions, and the description of details and assumptions embed in the codes I developed.

Partial results of the presented work have been submitted at ?.

### 3.1 Introduction to probability theory.

Uncertainty and probability are closely entangled. Anything we measure has an associated uncertainty, otherwise is not a complete measurement <sup>1</sup>. The term uncertainty must not be confused with the term error, which refers to the difference between the measured value of the quantity and the *true* value<sup>2</sup> of it (for Guides in

---

<sup>1</sup>Upper and lower limits are examples of incomplete measurements.

<sup>2</sup>The true value is that which ideally results when the uncertainty tends to zero.

Metrology 2008). It is commonly agreed that uncertainty of a measurement can be expressed in a probabilistic basis (for Guides in Metrology 2008). This means that whenever we measure a quantity, let's say  $a$ , the distribution of the repeated measurements of  $a$ , is a probability distribution function,  $p(a)$ . As any other probability distribution,  $p(a)$  satisfies the following properties:

**Property 1** It has units, those of the inverse of  $a$ .

**Property 2**  $p(a) \geq 0$ .

**Property 3**  $1 = \int_a p(a) da$ .

These properties hold regardless of the dimension of  $a$ , it means that the joint uncertainty of all measured quantities of an object is also a probability distribution. Furthermore, they also hold if the probability distribution is conditioned in any other quantity. Let's imagine that we measure the positions, projected in the plane of sky (the plane perpendicular to the line of sight), of one star, these measurements are conditioned in the magnitude (brightness) of the object we measure. If the object is too bright, like the sun, it will saturate the detector and it will render the measurement useless. On the other hand, if the object is too faint we simply will not have enough photons to measure it. So, the stellar positions in the sky, which we can call  $a$  and  $b$  because they are two dimensions, are conditioned on the magnitude,  $c$ , of the object. Therefore,  $p(a, b|c)$  must also satisfy:

- It has units of  $a^{-1}b^{-1}$ .
- $p(a, b|c) \geq 0$ .
- $1 = \int_a \int_b p(a, b|c) da \cdot db$ .

The link between joint and conditioned probabilities is given by the following symmetric definition:

$$\begin{aligned} p(a, b) &= p(a|b) \cdot p(b). \\ p(a, b) &= p(b|a) \cdot p(a). \end{aligned} \tag{3.1}$$

This can be further conditioned on  $c$  to obtain:

$$\begin{aligned} p(a, b|c) &= p(a|b, c) \cdot p(b|c), \\ p(a, b|c) &= p(b|a, c) \cdot p(a|c), \end{aligned} \tag{3.2}$$

If the joint probability of  $a$  and  $b$  can be factorised, this is

$$\begin{aligned} p(a, b) &= p(a) \cdot p(b), \\ p(a, b) &= p(b) \cdot p(a), \end{aligned} \tag{3.3}$$

then  $a$  and  $b$  are say to be *independent*. An alternative option is to say that  $a$  and  $b$  are *independent*, if the conditional probability of  $a$  on  $b$  is  $p(a|b) = p(a)$ .

The most important thing we can do with probability distributions is to integrate them. **Property 3** establish that the amount<sup>3</sup> of probability  $p(a)$  spread over the volume of the support of  $a$  adds to one. This Property allows us to *marginalise* any non-desired variable. Lets imagine again that  $a$  and  $b$  are the measured positions of some star and we have several measurements of these positions. Then we will have the joint probability distribution of them,  $p(a, b)$  (must likely it will be a bivariate gaussian but that does not matter now). If we are interested lets say in the mean value of  $a$ , we first must get rid of  $b$ . For it, we *marginalise* out  $b$  in the following way,

$$p(a) = \int_b p(a, b) \cdot db. \tag{3.4}$$

Then we compute the *expected value* of  $a$ ,  $E(a)$ , which is identified with the mean of  $a$  once we have drawn many realisations from its probability distribution. To compute it, we add all the possible values of  $a$  weighted by their probability. This is,

$$E(a) = \int_a a \cdot p(a) \cdot da. \tag{3.5}$$

---

<sup>3</sup>Which could be infinite, like in Dirac's delta.

Once again, these last two equations (3.4 and 3.5) hold in case they are conditioned in any other measurement. For example, the magnitude of the object, as in our previous analogy. Notice however that, once the brightness of the object lay within in the dynamic range of the detector,  $a$  and  $b$  became *independent* of the magnitude.

It is important to recall that the term measurement, and its unavoidable uncertainty, refer not just to directly measured quantities, like the photons (counts) and pixels in a CCD, but also to indirect measurements. Stellar magnitudes and positions in the sky, for example, are indirect measurements derived from the direct measurement of photons, pixels and telescope arrangement. This generalisation also applies to the measurement of parameters in any physical or statistical model, like the one I will describe in the following Section.

### 3.1.1 Bayes theorem

The definition of conditioned probability (Eq. 3.2) leads to the Bayes' theorem:

$$p(a|b, c) = \frac{p(b|a, c) \cdot p(a|c)}{p(b|c)}. \quad (3.6)$$

Integrating on  $a$  we find that,

$$\begin{aligned} p(b|c) \cdot \int_a p(a|b, c) \cdot da &= \int_a p(b|a, c) \cdot p(a|c) \cdot da \\ Z \equiv p(b|c) &= \int_a p(b|a, c) \cdot p(a|c) \cdot da. \end{aligned} \quad (3.7)$$

In this last equation  $Z$  refers to what is known as the *evidence*. I will come back to this *evidence* in a few paragraphs. This last Eq. also illustrates that  $p(b|c)$  is a normalisation constant which can be evaluated once  $p(b|a, c)$  and  $p(a|c)$  are known. These two terms are commonly referred as the *likelihood* ( $p(b|a, c)$ ), and the *prior* ( $p(a|c)$ ). Also the term the term  $p(a|b, c)$  is called the *posterior*. These names arise in the context of parametric inference, as I will describe in the next paragraph. However, it worths mention that, although formally the likelihood and the prior are probability distributions in  $b$  and  $a$ , respectively, for  $p(a|b, c)$  to be a probability distribution on  $a$ , it only suffices that the product of the likelihood times the prior

does not vanish everywhere or be negative anywhere<sup>4</sup>. In this case, they are called *improper* priors or likelihoods. If their product vanishes everywhere, which may be the case if the prior is terribly specified or if the likelihood does not take proper account of extreme data, then the posterior is not a probability distribution due to a division by zero. In any case, it makes no sense try to estimate the parameters of a model with zero evidence.

### 3.1.1.1 Models and parametric inference

In a broad sense, models are representation or abstraction of the knowledge about something. Sometimes the knowledge is shared by others, some time it is not. They are everywhere in our daily life: from the words we spoke every day, to the evolution of the species and the general relativity; from a kid's draw to the cosmological models. In science, however, we restrict the concept of model to a mathematical representation of the relations among the variables. If the model is parametric, the variables include the data  $\mathbf{D}$ , which the model attempts to model, and the parameters  $\theta$ . Parameters are free variables that allow the model to describe the data. Thus a parametric model  $\mathcal{M}$ , can be represented as:

$$\mathcal{M} = \{f(\mathbf{D}, \theta), \theta \in \Theta\}, \quad (3.8)$$

where  $f$  is the function that relates data and parameters, and  $\theta \in \mathbb{R}^k$  with  $k$  the dimension of  $\theta$ . Parametric inference is then the act of finding the distribution of  $\theta$  given the data  $\mathbf{D}$ .

The Bayes' theorem allows to perform parametric inference, which is called then Bayesian inference. In this context the Bayes' theorem is:

$$p(\theta|\mathbf{D}, M) = \frac{p(\mathbf{D}|\theta, M) \cdot p(\theta|M)}{p(\mathbf{D}|M)}. \quad (3.9)$$

where  $\theta$ ,  $\mathbf{D}$  and  $M$  are correspond, respectively, to the parameters in the model, the data which the model tries to describe, and the prior information used in the construction of the model. Whenever we have a model, we have prior knowledge over it. Actually, it can be classified in two kinds of prior information. One

---

<sup>4</sup>Although negative probabilities may have sense in quantum mechanics. See for example Dirac (1942)

refers to the prior information conveyed in the model, which I call  $M$ . This is the information that the creator of the model uses to establish the relations among the elements of the model: variables. The second kind of prior,  $p(\theta|M)$  refers to the statement the user of the model made of his/her believes about the probability distribution of the parameter values. This is indeed subjective. However, it is, in my opinion, less subjective than the former,  $M$ , prior information. At least in this last kind, the subjectivity is expressed objectively in a probabilistic, and therefore measurable way.

The likelihood of the data  $p(\mathbf{D}|\theta, M)$ , is a probability distribution on the data,  $\mathbf{D}$ . However, it is a function on the parameters,  $\theta$ , which corresponds to the function  $f$  of Eq. 3.8. This function is not necessarily a probability distribution on the parameters.

Almost always we assume that the distribution of measurements of one object is independent from the measured value of another object. It means that the collection of measurements we made about for example, the length of a pencil, is independent of the measured length of a pen. This assumption, however, does not always holds. Imagine for example, that we were conducting a statistical analysis on the length of the objects used to define our unit of length. In this case, the statistical distribution of any object is conditioned on the unit of measurement, which in turn is conditioned on the distribution of measurements of the rest of the objects. Therefore, the probability distributions of objects are not independent of themselves<sup>5</sup>. That said, if the data are independent from each other, then

$$p(\mathbf{D}) = \prod_{n=1}^N p(d_n), \quad n = \{1, 2, \dots, N\} \quad (3.10)$$

with  $N$  the number of measurements, and  $d_n$  the measurements of a single object. In this case, the joint probability  $p(\mathbf{D}|\theta, M)$  can be expressed as,

$$p(\mathbf{D}|\theta, M) = \prod_{n=1}^N p(d_n|\theta, M), \quad n = \{1, 2, \dots, N\}. \quad (3.11)$$

---

<sup>5</sup>This happens for example in satellite surveys for which their system of reference is defined by their own measurements. To avoid this issue their reference systems are anchored on independent measurements.



The term  $p(d_n|\theta, M)$  is the likelihood of datum  $d_n$ . This term also as called the *generative* model, since it contains the necessary information to generate the data<sup>6</sup>.

I interpret the Bayes' theorem, as the probabilistic way to update knowledge. To me, this relation embodies the process of knowledge improvement once we recognise that knowledge is uncertain. In my perspective, knowledge is always uncertain, even if its uncertainty is negligible given the current evidence that supports it. The Bayes' theorem helps us to update our prior knowledge by means of the data (once we multiply them by the likelihood). Then, the posterior probabilities, became our new knowledge. Furthermore, the Bayes' theorem also provides the objective way to compare two models or hypothesis, and update the prior information,  $M$ , used to construct them. This is called model selection, which I explain briefly in the next section.

### 3.1.2 Model Selection

Whenever we have a data set and two or more models that attempt to describe this data, the most straightforward thing to do is to compare these models. Almost always, we want to select the *best* model. Obviously the term *best* depends on the objective of research. For example, lets imagine that our data set consists of a set of bivariate points, for example the measurements of the positions of an object as function of time. If we were interested in reproducing exactly the same points in the data set, the *best* model will be a polynomial with degree equal to the number of points. This polynomial will pass trough all these points. Once we recognise the unavoidable uncertainty of the data, we realise that an exact representation of the data is of poor use. It fits also the noise of the data.

In general, we are interested in the predictive capabilities of a model, its ability to predict future observations rather than to replicate the ones we have. Thus, an exact representation of the observed data (an over-fitted model as in the previous example), will poorly describe any new data set. In this sense, an over-fitted model *memorises* the data rather than *learns* from them.

A model that *learns* from the data is that which obtains the *true* underlying

---

<sup>6</sup>Actually the generative model of the *true* data. To generate the observed data the noise process must be also specified.

relation embedded in the data. This *true* underlying relation produces the *true* data. The observed data results once the uncertainty is added to it. Nevertheless, we still need to select among different learning models.

We can draw some help from the commonly known Ockham’s razor or principle <sup>7</sup>. It says:

*Among competing hypotheses, the one with the fewest assumptions should be selected.*

Identifying hypothesis with models, this principle tells us we should choose the model with the fewest assumptions. I classify the assumptions of a model in two groups: fixed and free ones. The fixed assumptions belong to what I previously described as the prior information,  $M$ , used to construct the model. They may render the model more interpretable in the physically or statistically sense, or even give it coherency within the corpus of a theory. The free assumptions correspond to the parameters of the model. They give it more flexibility to fit the data, although they can also introduce degeneracy in the parametric space. For example, in the case of a straight line model, a fixed assumption is that the data is linearly related, whereas the free assumptions correspond to the slope and ordinate at the origin. Comparing a linear model to a quadratic one in which the constant term has been fixed, we see that they have the same number of free parameters but clearly the second one has an extra fixed assumption. Therefore, choosing the model with fewer free parameters does not necessarily means choosing the model with the fewest assumptions.

One of the great advantages of the Bayesian methodology is that it incorporates directly the Ockham’s principle. Suppose that we want to compare two models,  $M_1$  and  $M_2$ , which we assume describe the data set  $\mathbf{D}$ . Each model has prior probabilities,  $p(M_k)$  and likelihoods  $p(\mathbf{D}|M_k)$  (with  $k = 1, 2$ ). Notice that now, I use the Bayes’ theorem for models and not to parameters within a model, as before. So, the prior probabilities of the models reflect our believes about the fixed assumptions within each model. On the other hand, the likelihood of the data given the model, is related to the parameters (the free assumptions) and priors, both within a model. This likelihood of the data given the model corresponds to the

---

<sup>7</sup>The origin of this motto and its exact phrasing is beyond the scope of this work. I just mention that paradoxically, an ancient formulation is attributed to Ptolomey: "We consider it a good principle to explain the phenomena by the simplest hypothesis possible" (Franklin 2002)

evidence of the model, Eq. 3.7. This evidence in terms of the model parameters,  $\theta_k$ , is now

$$p(\mathbf{D}|M_k) = \int_{\theta_k} p(\mathbf{D}|\theta_k, M_k) \cdot p(\theta_k|M_k) \cdot d\theta_k. \quad (3.12)$$

The Bayes' theorem applied to models instead of individual parameters tells us that

$$p(M_k|\mathbf{D}) = \frac{p(\mathbf{D}|M_k) \cdot p(M_k)}{p(\mathbf{D})}. \quad (3.13)$$

with  $k = 1, 2$ . Since there are only two models, their prior probabilities are related by  $p(M_1) = 1 - p(M_2)$ . Therefore,

$$p(M_k|\mathbf{D}) = \frac{p(\mathbf{D}|M_k) \cdot p(M_k)}{p(\mathbf{D}|M_1) \cdot p(M_1) + p(\mathbf{D}|M_2) \cdot p(M_2)}. \quad (3.14)$$

From this last Equation, the ratio of the posterior distributions is:

$$\frac{p(M_1|\mathbf{D})}{p(M_2|\mathbf{D})} = \frac{p(\mathbf{D}|M_1) \cdot p(M_1)}{p(\mathbf{D}|M_2) \cdot p(M_2)}. \quad (3.15)$$

This ratio provides an objective measure of how better model  $M_1$  is when compared to model  $M_2$ , under the measure provided by the data  $\mathbf{D}$  by means of the evidence. When both prior probabilities  $p(M_1)$  and  $p(M_2)$  are set equal, the ratio of posteriors equal the ratio of likelihoods. This is known as the *Bayes factor* (for a similar derivation and some examples of its application see Kass & Raftery 1995). Even in the equal priors case, the evidences themselves, Eq. 3.12, embody the Ockham's principle. Indeed, the evidence is a measure of the prior times the likelihood, this time for parameters in a single model. The larger the number of parameters (assumptions), the larger the volume, in parametric space, over which the likelihood of the data spreads. Since the likelihood is not a probability distribution on the parameters, it does not integrate to one, even if the priors are uniform. The evidence also penalises the assumptions made in the priors of the parameters. The most concentrated the prior is, the less of the likelihood contributes to the evidence.

Thus, the Bayes' theorem is the way to update knowledge, either if it refers to models or to parameters within a model.

### 3.1.3 Membership probability

In the previous Section, I derived, by means of the Bayes' theorem, the probability of models  $M_1$  and  $M_2$  given the data  $\mathbf{D}$ . Now, I describe the same problem but instead of the likelihood of a data set I do it for a single datum. This is, the probability of model  $M_1$  or  $M_2$ , given the datum  $\mathbf{d}$ . This is known as the membership probability of the datum  $\mathbf{d}$  to model or class,  $M_k$  ( $k = 1, 2$ ). The Bayes' theorem in this case is,

$$p(M_k|\mathbf{d}) = \frac{p(\mathbf{d}|M_k) \cdot p(M_k)}{\sum_{k=1}^2 p(\mathbf{d}|M_k) \cdot p(M_k)} \quad (3.16)$$

## 3.2 Bayesian hierarchical Models

### 3.2.1 Generalities

Bayesian formalism requires the establishment of priors. As mentioned before, priors represent the *a priori* believe the user of the model has about the possible values that parameters of the model can take. This is indeed subjective. This subjectivity is the main source of criticism from the non-bayesian community <sup>8</sup>.

Bayesian hierarchical models, in the following (BHM) can be grouped into the Empirical Bayes methods. In these later ones, the prior distributions are inferred from the data, rather than being directly specified as in common Bayesian methods. In BHM, priors are specified by parametric distributions whose parameters (called hyper-parameters) are also drawn from a parametric distribution in a hierarchical fashion. For this reason, hierarchical models are also called multilevel models. A fully-BHM is that in which the parameters at its higher hierarchy are drawn from a non-parametric distribution. Given its properties, BHM represent the most objective way to the establishment of prior distributions (Gelman 2006). Despite the possible high hierarchy of a BHM, for it to be valid, the class of prior distribution must allow the *true* value of its parameter (Morris 1983). For this reason, the updating of knowledge is an important step in the any Bayesian study. If the posterior distribution is in total discrepancy with the prior distribution, or even worst, when the posterior is not fully allowed by the prior distribution (as in

---

<sup>8</sup>See Gelman (2012) for a discussion on the ethical use of prior information

a truncated prior for example), then we must update our prior and allow the data to be fully expressed. Otherwise, the posterior could be biased.

Despite its theoretical advantages, BHM are difficult to evaluate since they require far more parameters than standard Bayes methods. Furthermore, their hierarchy (levels) must stop at some point. There are at least two approaches to stop this hierarchy. The first one is to use a non parametric distribution for the hyper-parameters at the higher level. This renders, as previously said, a fully-BHM. However, to use a non-parametric distribution we must have certain prior knowledge about it, which, most of the time is not the case. Another more practical alternative is to give a point estimate, usually the mean or the mode, for the distribution of the hyper-parameter at the top of the hierarchy.

Although in BHM the parameters values of the prior distributions are inferred from data, the user of the model has the important task of specifying the kind of distribution to be used for the prior. Selecting the kind of prior distributions continues to be an active area of research. Common options are include conjugate, non-informative, and weakly informative priors. Conjugate priors are those in which the posterior distribution turns out to be in the same family as the prior distribution, they are called the conjugate of the likelihood. Non-informative and weakly informative priors, as they names indicate, provide intentionally weaker information or no information at all for the prior. Weakly informative priors are the recommended ones (see for example the works of Gelman 2006; Huang & Wand 2013; Chung et al. 2015a). Despite the kind of prior distribution chosen, we must always evaluate the prior distribution in terms of the posterior, and check if this last one make sense (Gelman 2006; Gelman et al. 2013, Chap. 6).

### 3.2.2 Examples

Since BHM usually need more parameters than standard techniques, it restricted its use until modern computers were widely available. Although the idea of BHM was already present in the 1960s, its application to inference of normal distributions and linear models appears in the 1970s (see Good 1980, for an historical perspective of BHM). In modern days, BHM have a wide range of applications. Some examples of its application are in Gelman & Hill (2007) for the social sciences, Fei-Fei & Perona (2005) for vision recognition and, Diard & Bessiere (2008)

for robot navigation.

BHM are widely applied in astrophysics. Although, originally its applications were use mainly in the domain of cosmological parameters inference (see for example the works of Feeney et al. 2013; March et al. 2014; Anderes et al. 2015; Shariff et al. 2016; Alsing et al. 2017), its use was adopted in other domains. Some examples include the study of: the eccentricity distribution of binary stars Hogg et al. (2010b), the Cepheids (Barnes et al. 2004) and RR Lyrae distances (Jefferys et al. 2007), the chemical composition (Wolfgang & Lopez 2015) and albedos of exoplanets (Demory 2014), extinction maps (Sale 2012), stellar parameters (Shkedy et al. 2007), and the present day mass function (Tapiador et al. 2017).

### 3.2.3 Graphical representation.

Due to the generally large number of parameters in BHM, its interpretation benefits from a graphical representation. Probabilistically Graphical Models (PGM) are graphs that depict the conditional relation among the elements in a model. The elements in a probabilistic model could be constants or stochastic variables. The conditional relations that link elements could be deterministic or stochastic.

In PGM, stochastic variables are represented with circles while constants with squares. If the variable is known, as in the case of the data, it is represented with a filled symbol, otherwise with an empty symbol. Stochastic relations are depicted with solid lines while deterministic ones with dashed lines. If there is no line between two given variables, it indicates that they are assumed to be independent. Variables that repeat together, as in the case of the data, are grouped within a plate. The number of repetitions is indicated in one corner of the plate. The community generally agrees in these set of standard representations (for more details on PGM see for example the book of Koller & Friedman 2009). Figure 3.1 shows a simple example of a PGM for the inference of the parameters of a normal distribution.

Figure 3.1: PGM representing the parametric inference of a normal distribution.

### 3.3 Modelling the data

Creating a model is a complex task. As previously mentioned, a model is the mathematical representation of the knowledge about something. In this work my objective is the modelling of the DANCe data related to NYOC (nearby young open clusters). Since my aim is the statistical description of the NYOC population, most of the time, the relations I use are statistical.

Modelling a data set demands the gathering and sorting of the prior knowledge. This last refers to the knowledge about the data set, the object of study, the statistical techniques that may help to attain the objective, and the computational resources at hand. I collected this knowledge from three main sources: the standard references (e.g. articles and books), my colleagues and experts (*knowledge elicitation*), and my self. Arraigning the prior knowledge into the model is an iterative and therefore continuous process. In this Section I describe an snapshot of this process. The state of the model once the article ? was submitted. Later in Section 3.6 I will give a brief description of the model development process in the context of its coded versions.

In this Section I will describe one crucial aspect of the DANCe DR2 data set: the missing values. Then I will describe the relevant knowledge of NYOC and how I embedded this knowledge in the the data model in terms of field and cluster models. Finally, I will describe the details of the prior distributions and of its hierarchy.

In the following, whenever I use the pronoun *we*, it refers to the authors of ?, where a synthesis of this work is presented. I use the pronoun *I* to emphasis that the particular idea or task was done by me.

#### 3.3.1 Missing values

Missing values can happen due to different processes. From the physical perspective, they can arise due to faint or bright sources that produce counts which are outside the dynamical range of the detector. They can also emerge due to for example detector or random issues (e.g. electronic failures or cosmic rays). From the statistical perspective however, the most important aspects of missing values are the probability distribution of their occurrence, and the fact that they are par-

tially or completely missing. A partially missing value (or partially observed) is that for which an upper or lower limit is given whereas a completely missing value is simply not available at all. The DANCe survey contains only (so far) completely missing values. Upper and lower limits could also be inferred from the data provided that missing values occur outside these limits. However, I leave aside this task since missing values in DANCe data occur also at the interior of the variables domain.

In terms of probability, there is no distinction between missing values and parameters. Therefore, we can marginalise missing values as we do with any other nuisance parameter. If datum  $\mathbf{d}$  has a missing entry,  $\mathbf{d} = \{d_1, d_2, \dots, \text{mis}, \dots, d_n\}$ , with  $n$  the dimension of  $\mathbf{d}$ , then, the likelihood of this datum, given model parameters  $\theta$  is

$$p(\mathbf{d}|\theta) = \int_{-\infty}^{\infty} p(\{d_1, d_2, \dots, \text{mis}, \dots, d_n\}|\theta) d\text{mis}. \quad (3.17)$$

Throughout this work, missing values are marginalised in this way.

I made a remark of a point that may seem obvious but it is important to remember. Let  $p(a)$  be a probability distribution,  $A$  a random sample of  $n$  point from it, and  $p_A(a)$  the empirical probability distribution of  $A$ . Let  $B$  be a non-random sample of  $p(a)$  with  $n$  elements, and empirical probability distribution  $p_B(a)$ . In the limit of  $n \rightarrow \infty$ ,  $p_A(a) = p(a)$  however  $p_B(a) \neq p(a)$ . Therefore,  $p_A(a) \neq p_B(a)$ . Similarly, in data with missing values, if the missing value pattern is not random, the distribution of the completely observed data (with non-missing values) differs from that of all the data. In ? we show this subtle but important difference in the case of the DANCe data set.

### 3.3.2 The generative model

Since the objective of this work is the statistical study of NYOC we must separate them from the field population. To perform this separation, we use the data, which as always, is uncertain. Therefore, the separation is also uncertain. As mentioned in Chapter ??, given the current set of variables, the cluster and the field are entangled. To probabilistically disentangle them, we must provide probabilistic models for both populations. These models are the likelihoods. With



these likelihoods and prior probabilities for the cluster and field model, we are able to compute the membership probability (Eq. 3.16) of each object in our data set to the cluster model. This model is assumed to represent the cluster population. First, this model must be learnt from the data.

The learning process demands a set of  $N$  binary integers  $\mathbf{q}$ , one  $q_n$  for each object. Each of these two possible values represent one of the two mutually exclusive possibilities: the object belongs to the cluster ( $q_n = 1$ ) or to the field population ( $q_n = 0$ ). Let  $\theta$  and  $p_c$  be the parameters and model of the cluster. Also,  $\phi$  and  $p_f$  the parameters and model of the field. Then, the likelihood of the data is,

$$p(\mathbf{D}|\mathbf{q}, \theta, \phi) = \prod_{n=1}^N p_c(\mathbf{d}_n|\theta)^{q_n} \cdot p_f(\mathbf{d}_n|\phi)^{(1-q_n)}. \quad (3.18)$$

This  $\mathbf{q}$  is now marginalised using a probability for it, a prior probability or measure which is set in terms of a new and unique parameter  $\pi$ , which represent the *prior* probability that an object belongs to the field. Thus, the probability of  $\mathbf{q}$  is

$$p(\mathbf{q}|\pi) = \prod_{n=1}^N (1 - \pi)^{q_n} \cdot \pi^{(1-q_n)}. \quad (3.19)$$

and the marginalisation runs as

$$\begin{aligned} p(\mathbf{D}|\pi, \theta, \phi) &= \int_{\mathbf{q}} p(\mathbf{D}, \mathbf{q}|\pi, \theta, \phi) \cdot d\mathbf{q} \\ &= \int_{\mathbf{q}} p(\mathbf{D}|\mathbf{q}, \pi, \theta, \phi) \cdot p(\mathbf{q}|\pi) \cdot d\mathbf{q} \\ &= \int_{\mathbf{q}} \prod_{n=1}^N p_c(\mathbf{d}_n|\theta)^{q_n} \cdot p_f(\mathbf{d}_n|\phi)^{(1-q_n)} \cdot \prod_{n=1}^N (1 - \pi)^{q_n} \cdot \pi^{(1-q_n)} \cdot d\mathbf{q} \\ &= \int_{\mathbf{q}} \prod_{n=1}^N [(1 - \pi) \cdot p_c(\mathbf{d}_n|\theta)]^{q_n} \cdot [\pi \cdot p_f(\mathbf{d}_n|\phi)]^{(1-q_n)} \cdot d\mathbf{q} \\ &= \prod_{n=1}^N (1 - \pi) \cdot p_c(\mathbf{d}_n|\theta) + \pi \cdot p_f(\mathbf{d}_n|\phi). \end{aligned} \quad (3.20)$$

This last equality is a rather complicated derivation which can be found on Press

(1997); Hogg et al. (2010a) for  $p_c$  and  $p_f$  in the exponential family. Also, a general derivation of this expression is given by Jaynes (2003). He obtains it assuming individual unknown probabilities  $p_n$  instead of  $q_n$  and marginalising over them by the aid of a prior.

Thus, the *generative model* or likelihood of the datum  $\mathbf{d}_n$  is

$$p(\mathbf{d}_n|\pi, \boldsymbol{\theta}_c, \boldsymbol{\theta}_f, \mathbf{u}_n) = \pi \cdot p_f(\mathbf{d}_n|\boldsymbol{\theta}_f, \mathbf{u}_n) + (1 - \pi) \cdot p_c(\mathbf{d}_n|\boldsymbol{\theta}_c, \mathbf{u}_n), \quad (3.21)$$

where  $\boldsymbol{\theta}_f$  and  $\boldsymbol{\theta}_c$  indicate the cluster and field parameters, while  $\mathbf{u}_n$  refers to the datum uncertainty. The probabilities  $p_f(\mathbf{d}_n|\boldsymbol{\theta}_f, \mathbf{u}_n)$  and  $p_c(\mathbf{d}_n|\boldsymbol{\theta}_c, \mathbf{u}_n)$  are the field and cluster models, respectively. These models are explained in detail in the next two sections.

In the following, I assume that the observed quantities, even if they contain missing values, resulted from the convolution<sup>9</sup> of the *true* quantities with a source of uncertainty. If I were to assume that uncertainties of individual objects were drawn from the same uncertainty distribution, then these objects will be independent and identically distributed (commonly known as i.i.d). However, this assumption is not necessary. Instead, we model the data with its intrinsic heteroscedasticity. It means that individual observations have different dispersions. However, we assume that the multivariate normal is the family distribution for these uncertainties. This assumption is standard practice and is also supported by the large and heterogeneous origins of the DANCe data.

### 3.3.3 The field population

To model the field population, we assume that the joint probability distribution of the data can be factorised into the probability distributions of proper motion and photometry. Thus, they are independent, at least conditioned on the parameters. It is also assumed that both distributions are described by Gaussian Mixture Models (GMM). The flexibility of GMM to fit a variety of probability distribution geometries make them a suitable model to describe the density of the

---

<sup>9</sup>The addition of two stochastic variables is analogous to the convolution of their probability distributions.

heterogeneous data from the DANCe DR2.

A GMM is a probability distribution resulting from the linear combination of  $M$  gaussian distributions,  $\pi_m \cdot \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ , with  $m = 1, 2, \dots, M$ . Where  $\pi_m$  are the amplitudes or fractions, which must add to one. Thus

$$p_{GMM}(x|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{m=1}^M \pi_m \cdot \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \quad (3.22)$$

According to Bouy et al. (2015b), the number of Pleiades candidate members in their data set is 2109. This means that the number of field objects dominates. Thus, it can be assumed that any reclassification of candidate members will have a negligible impact on this figures. Therefore, it seems reasonable to assume that the GMM describing the field population can be fixed during the process of cluster parameters inference. Let me elaborate on this assumption.

The field objects of Bouy et al. (2015b) are those whose cluster membership probability is lower than 0.75. These are  $\approx 98,000$  objects. These authors selected this probability threshold based on the analysis that Sarro et al. (2014) made of their methodology when applied to synthetic data. Sarro et al. (2014) report, at probability threshold  $p = 0.75$  contamination and true positive rates of  $\approx 8\%$  and  $\approx 96\%$  respectively. Therefore the number of hypothetically misclassified objects, 12% of the cluster members,  $\approx 258$ , is negligible compared to the size of the data set (100,000 objects, the restricted sample). It represents a negligible fraction ( $\lesssim 0.26\%$ ). Furthermore, under the assumption that the work of Sarro et al. (2014) is correct, the miss classified objects concentrate on cluster membership probabilities near the classification threshold. Therefore, they will also group in an area of the physical "boundary" between the cluster and the field (I call it physical because it is on the observable variables and not in the probability threshold)). This is the area of highest entanglement. It does not meant that misclassified objects will not lay in the core of the cluster (objects with high membership probabilities). It means that the occurrence of these cases will be lower. This physical "boundary" will correspond, in proper motions space to a halo around the cluster centre. In photometry, however, this boundary will run all along the cluster sequence in the CMDs. All previous assumption are there to justify that the negligible fraction of hypothetical misclassified is not concentrated in the physical space.

Therefore, if the misclassified objects are a few and spread over the space, their contribution to the parameters of the GMM describing the field population can be neglected. Thus, the parameters of the GMM remain fixed and out of the inference process.

The number of gaussians in each GMM was found using the Bayesian Information criterion (BIC, Schwarz 1978). The BIC is a model selection criteria that aims at avoiding over fitting. It represents a compromise between the likelihood,  $\mathcal{L}$ , of the  $n$  data points, and the number of parameters,  $k$ . This is,

$$BIC = \ln n \cdot k - 2 \ln \mathcal{L} \quad (3.23)$$

To estimate the parameters of the GMM that maximise the likelihood, we used the Expectation Maximisation (EM) algorithm. However, the missing values in the photometry prevent the use of the standard form of the algorithm (see for example Chapter 9 of Bishop 2006). Instead, the parameters were estimated with the modified version of the EM for GMM of McMichael (1996). On it, objects with missing values also contribute to estimate the maximum-likelihood (ML) parameters. The number of gaussians suggested by the BIC for this mixture is 14.

Regarding the proper motions, the standard EM for GMM can be used. However, the BIC favours models in which the number of gaussians is large, their fractions small and their variances also large. To circumvent this issue, a uniform distribution was added to the GMM. The EM algorithm was modified accordingly. The BIC applied to this new mixture of distributions renders reasonable results. This modification improves the likelihood while reduces the number of parameters. The number of gaussians suggested by the BIC for this mixture is 7, plus the uniform distribution.

Thus, the field likelihood  $p_f(\mathbf{d}|\boldsymbol{\theta}_f, \mathbf{u})$  of an object with measurements  $\mathbf{d}$ , given

parameters,  $\theta_f$ , and standard uncertainties  $\mathbf{u}$  is

$$p_f(\mathbf{d}|\theta_f, \mathbf{u}) = \left[ c \cdot \pi_{f,pm,0} + \sum_{i=1}^7 \pi_{f,pm,i} \cdot \mathcal{N}(\mathbf{d}_{pm} | \boldsymbol{\mu}_{f,pm,i}, \boldsymbol{\Sigma}_{f,pm,i} + \mathbf{u}_{pm}) \right] \cdot \left[ \sum_{i=1}^{14} \pi_{f,ph,i} \cdot \mathcal{N}(\mathbf{d}_{ph} | \boldsymbol{\mu}_{f,ph,i}, \boldsymbol{\Sigma}_{f,ph,i} + \mathbf{u}_{ph}) \right]. \quad (3.24)$$

In this equation,  $\theta_f$  refers to the set of field parameters,  $\boldsymbol{\pi}_f, \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f$ , which are, respectively, the fractions, means and covariance matrices of the GMM. The first and second brackets represent the proper motion and photometric models, respectively. The first term of the proper motion model is the uniform distribution. In it,  $c$  is a constant determined by the inverse of the product of the proper motions ranges (see Table 2.1), and  $\pi_{f,pm,0}$  is the fraction of this uniform distribution. The second term in the same bracket is the mixture of gaussians with means  $\boldsymbol{\mu}_{f,pm}$  and covariance matrices  $\boldsymbol{\Sigma}_{f,pm} + \mathbf{u}_{pm}$ .

I refer the interested reader to Appendix ??, which contains specific details of the GMM presented in this section.

### 3.3.4 The cluster population

Similarly to what it was assumed for the field population, for the cluster population we assumed also that its data distribution can be factorised in the product of the proper motions distributions times the photometric distribution. It is known that unresolved systems of stars (groups of stars that, given the spatial resolution of the telescope, are seen as an individual object) have an increased brightness proportional to the multiplicity of the system. In particular, if an unresolved system is made of two equally luminous objects, then its magnitude is 0.752 times brighter than that of an individual object. This is the case of an equal mass binary.

Sarro et al. (2014) show evidence of an equal-mass binaries (EMB) sequence in the Pleiades cluster. Those authors model this parallel sequence assuming that the number of objects in this sequence is 20% of the total number of members. In this work, we also model this EMB sequence. However, we do not assume its proportion and rather infer it from the data.

Unresolved multiple systems have an impact on proper motion. From an astrophysical point of view, in stellar clusters, massive objects fall towards the centre of the gravitational potential in a higher rate than less massive ones due to stellar encounters (reference needed). From an astronomical point of view, an unresolved binary system shifts the photo centre of its images when compared to that of a single object. For the previous reasons, we model the EMB as an independent population in the proper motions, and pair this model to its photometric parallel 0.75 displaced model. This statistical model allows a comparison of the EMB population with that of the rest of the cluster. We call all non EMB objects single stars. Although this is an abuse of the terminology because there are binaries and multiple systems with different mass ratios, it keeps the text more readable.

*Photometric model of equal-mass binaries and single stars*

To model the cluster sequence in the CMDs, both for single and EMB, we use cubic splines for each of the  $YJHK_s$  vs  $CI$  CMDs. This assumption roots on the fitting properties of splines. I tried several polynomial bases (Laguerre, Hermite, Chebyshev) but no matter the order, they were not able to fit the sequence, particularly in the region around  $CI \approx 3$ , where the slope is higher. Despite their flexibility, these splines require more parameters than common polynomials. If represented in terms of B-splines series, in addition to the coefficients of the series, they require the specification of points called knots. These knots represent the starting and ending points of the spline sections. By definition, a basis spline (B-spline) of order  $n$  is a piece wise polynomial function of order  $n$  in the interval  $t_0 \leq x \leq t_n$ . The boundary and internal points  $\mathbf{t}$  are called knots. In particular, for a given set of knots, there is only one B-spline, thus the name basis spline. Thus, any spline function of order  $n$  can be represented as a series of b-splines. In particular, any cubic spline can be represented as,

$$S_3(CI, \mathbf{t}) = \sum_i \beta_i \cdot B_{i,3}(CI, \mathbf{t}). \quad (3.25)$$

Where  $B_{i,n}$  are given by the Cox-de Boor recursive formula. For more details on splines and the Cox-de Boor formula see ?.

Despite their fitting properties, B-splines have an issue when inferring simultaneously their coefficients and knots: there is multi modality in the parametric

space (Lindstrom 1999). It means that at least more than one combination of parameters produces the same solution. To avoid this multi modality, I decided to keep the knots fixed throughout inference. This decision, although reduces the flexibility of the splines, allows a still better fit than common polynomials. To obtain the ML estimate of the knots I use the algorithm of Spiriti et al. (2013). This algorithm, implemented in the *freeknotsplines* R package, allows to simultaneously obtain the best truncation value for the spline series. It uses the BIC to select among different models. In order to obtain these figures, I used the candidate members of Bouy et al. (2015b). The BIC indicates that seven coefficients is the best number with knots at  $\mathbf{t} = \{0.8, 3.22, 3.22, 5.17, 8.0\}$ . I tested different number of knots, ranging from two to nine, with five the best configuration given by the BIC.

As I mentioned in the introduction to this Section, we assume that the observed photometric quantities are drawn from a probability distribution resulting from the convolution of the observed uncertainties, with the *true* quantities. Here comes an expiation. I recognise that the model is far from perfect and that there are several astrophysical phenomena that it could not address. Either because I do not have the knowledge or because they are too complicated to model. These phenomena include but are not limited to age, metallicity and distance dispersions, unresolved systems (other than EMB), variability, transits, etc. So, instead of modelling them I assume that all of them, even the unknown, contribute to an intrinsic photometric dispersion. Given the large and unknown sources contributing to this intrinsic dispersion, we can safely assume that it is gaussian, in fact multivariate gaussian. But most importantly, we distinguish this dispersion from the photometric uncertainty of individual measurements. If we were to assume no *true* intrinsic dispersion, then any deviation from the *true* quantities should have to be explained *only* by the observational uncertainty.

The splines model the mean of the *true* photometric quantities. They are the mean of the multivariate gaussian distribution that models the *true* photometry. The covariance matrix of this distribution represents the width of the intrinsic dispersion. These *true* photometric quantities and their mean are dictated by the parameters.

Since the splines are parametrised by the true *CI* of each object, we have more

parameters than objects <sup>10</sup>. This *true CI* is unknown even if its observed value is not missing. To solve this problem (it is a computational problem!) we marginalise these nuisance parameters (See Eq. eq:examplemarginal) . To marginalise these *CI* we need a measure or prior. We provide a prior, in a hierarchical way, (thus the name Bayesian Hierarchical model) and then marginalise the parameter. This marginalisation leaves behind a precise estimate of the parameters of the prior distribution. To this estimate all objects contributed, paradoxically also those with a missing *CI*. This is the force of the BHM.

We model the prior for the *true CI* distributions as a truncated ( $0.8 \leq CI \leq 8$ ) univariate GMM with five components whose parameters are also inferred from the data. We chose five components as BIC suggested. I applied BIC to the results found using the EM algorithm and the candidate members of Bouy et al. (2015b). I tested higher values but the posterior distribution did not changed much, thus I preferred the BIC value.

$$\begin{aligned} p(\mathbf{d}_{ph}|\boldsymbol{\theta}_c, \mathbf{u}_{ph}) &= \int p(\mathbf{d}_{ph}, CI|\boldsymbol{\theta}_c, \mathbf{u}_{ph}) \cdot dCI \\ &= \int p(\mathbf{d}_{ph}|CI, \boldsymbol{\theta}_c, \mathbf{u}_{ph}) \cdot p(CI|\boldsymbol{\theta}_c, \mathbf{u}_{ph}) \cdot dCI. \end{aligned} \quad (3.26)$$

In the previous equation,  $\mathbf{d}_{ph}$ ,  $\mathbf{u}_{ph}$  and,  $\boldsymbol{\theta}_c$  correspond to the photometric measurements, standard photometric uncertainties, and the cluster parameters, respectively. The term  $p(\mathbf{d}_{ph}|CI, \boldsymbol{\theta}_c, \mathbf{u}_{ph})$  corresponds to the multivariate gaussian associated with the intrinsic dispersion of the cluster. The *CI* dictates the *true* photometric quantities by means of the splines. The term  $p(CI|\boldsymbol{\theta}_c, \mathbf{u}_{ph})$  correspond to the truncated GMM which we use as a measure for the *true CI*. Appendix ?? contains more details on this marginalisation and the probability distribution involved on it.

We use the observed *CI* and magnitudes to reduce the computing time of the marginalisation integral by avoiding regions in which the argument is almost zero (i.e. far away from the measured values). The process is the following: first, we compare the observed photometry to the true one (i.e. the cluster sequence given by the splines) and find the closest point,  $p$ , using the Mahalanobis metric. This metric

---

<sup>10</sup>Although this sounds crazy, the rules of probability calculus do not discard this possibility.



uses the sum of the observational uncertainty with the intrinsic dispersion of the cluster sequence as covariance matrix. To define the limits of the marginalisation integral, we use a ball of 3.5 Mahalanobis distances around point  $p$ . Contributions outside this ball are negligible ( $< 4 \times 10^{-4}$ ).

Since we model the true photometric quantities of the equal-mass binaries with a parallel sequence displaced 0.75 magnitudes into the bright side (twice the luminosity implies an increase of 0.75 in magnitudes), the only extra parameter needed is the fraction of equal-mass binaries to the total of cluster members.

#### *Proper motion model of equal-mass binaries and single stars*

We model the proper motions of equal-mass binaries and single stars with a GMM whose parameters are inferred as part of the hierarchical model. The number of gaussians, however, remains fixed throughout inference. Following the BIC criterium, we select four and two gaussians for single and equal-mass binaries, respectively. Furthermore, we also assume that the gaussians in the proper motions GMM share the same mean, one for single stars and one for equal-mass binaries (which need not be equal).

## 3.4 Priors

In a Bayesian framework, each parameter in the generative model has a prior, even if it is uniform or improper. The priors we assume are intended to fall in the category of weakly informative priors. A weakly informative prior, following Gelman (2006), is that in which "the information it does provide is intentionally weaker than whatever actual prior knowledge is available". We chose this kind of priors because they show better properties regarding the regularisation and stability of the posterior computation (see Gelman et al. 2008; Chung et al. 2015b, for examples of this kind of priors). We group priors into three main categories, those for fractions, and those for parameters in the proper motion and in the photometrical models. Here we explain the kind of distributions we use for the priors. In Appendix ??, we give details on the particular parameter values we chose for these distributions.

Fractions are defined for mixtures, which can be GMM or the cluster-field mixture (Eq. 3.21), and quantify the contribution of each element to the mixture.

Thus, they must add to one and be bounded by the  $[0, 1]$  interval. For priors of fractions we use the multivariate generalisation of the beta distribution: the Dirichlet distribution.

We set the priors of means and covariance matrices in the proper motions GMM as bivariate normal and Half- $t$  distributions, respectively. See Huang & Wand (2013) for the advantageous properties of the Half- $t$  distribution.

Photometric priors include three categories, those concerning the *true CI*, the splines coefficients, and the cluster sequence intrinsic dispersion. For the priors of the means and variances of the true *CI* GMM, we use the uniform and Half-Cauchy distributions, respectively. For the coefficients in the spline series we set the priors as univariate normal distributions. Finally, we use the multivariate Half- $t$  distribution as a prior for the covariance matrix modelling the intrinsic dispersion of the cluster sequence.

### 3.5 Sampling the posterior distribution

There are three possible approaches to obtain the posterior distributions of the parameters in our model: an analytical solution, a grid in parameter space, and the Markov Chain Monte Carlo (MCMC) methods. Given the size of our data set ( $10^5$  objects) and the dimension of our model (85 parameters), the analytical solution and the grid approach are discarded a priori.

The MCMC methods offer a feasible alternative to this problem. Briefly, they consist of a particle (or particles) which iteratively moves in the parameter space. Among the many MCMC methods that exist, we select the *stretch* move which is an affine invariant scheme developed by Goodman & Weare (2010). It is implemented to work on parallel in the Python routine *emcee* (Foreman-Mackey et al. 2013). We chose *emcee* due to the following properties: i) the affine invariance allows a faster convergence over common and skewed distributions (see Goodman & Weare 2010; Foreman-Mackey et al. 2013, for detail), ii) the parallel computation distributes particles over nodes of a computer cluster and thus reduces considerably the computing time, and iii) it requires the hand-tuning of only two constants: the number of particles, and  $a$ , the parameter of *stretch* distribution (see Eq. 9 of Goodman & Weare 2010). We use 170 particles (twice the number of parameters)

and a value of  $a = 1.3$ . These keep the acceptance fraction in the range  $0.2 - 0.5$ , as suggested by Foreman-Mackey et al. (2013).

We use *CosmoHammer* (Akeret et al. 2013), a front-end of *emcee*, to control the input and output of data and parameters, as well as the hybrid parallel computing. We run it on a 80 CPUs (cores) computer cluster with 3.5 GHz processors. However, instead of using OpenMP as (Akeret et al. 2013) did, we use the *multi-processing* package of python to distribute the computing of the likelihood among cores in each cluster node.

Since the evaluation of the likelihood is computationally expensive (it takes approximately 30 days to run in the previously described computer cluster), we proceed similarly to Akeret et al. (2013). We provide *emcee* with an optimised set of values of the posterior distribution. These values can be thought of as a ball around the maximum-a-posteriori (MAP) solution. We find them with a modified version of the Charged Particle Swarm Optimiser (PSO) of Blackwell & Bentley (2002). It avoids the over-crowding of particles around local best values. The charged version retains the PSO exploratory property by repelling particles that come closer than a certain user specified distance to each other. The repelling force mimics an electrostatic force, thus the name charged PSO.

The modification that we introduce to the charged PSO relates only to the measuring of distance between particles. The algorithm of Blackwell & Bentley (2002) computes these distances in the entire parametric space. We find this approach unsuitable for our problem. In it, parameters have different length scales (for example, fractions and proper motions). Therefore, we measure distance between particles and apply the electrostatic force independently in each parameter. Thus, the electrostatic force comes into action only when the relative distance between particles is smaller than  $10^{-10}$ . We chose this value heuristically.

The PSO does not warrant the finding of the global maximum of the score function (see Blackwell & Bentley 2002; Clerc & Kennedy 2002, and references therein). Therefore, we iteratively run PSO and 50 iterations of *emcee* (with the same number of particles as the PSO) until the relative difference between means of consecutive iterations is lower than  $10^{-7}$ . The iterations of *emcee* guarantee the spreading of the PSO solution without losing the information gained. After convergence of the PSO-*emcee* scheme, we run *emcee* with 175 walkers, until con-

vergence. Neither scheme, PSO alone or PSO-*emcee*, guarantees to find the global maximum and their solution could be biased. However, we use them to obtain a fast estimate of the global maximum, or at least, of points in its vicinity. Nevertheless, the final *emcee* run, during the burning phase, erases any dependance on these initial solutions.

Convergence to the target distribution occurs when each parameter enters into the stationary equilibrium, or normal state. The Central Limit Theorem ensures that this state exists. See Roberts & Rosenthal (2004) for guaranteeing conditions and Goodman & Weare (2010) for *irreducibility* of the *emcee* stretch move. The stationary or normal state is reached when, in at least 95% of the iterations, the sample mean is bounded by two standard deviations of the sample, and the variance by the two standard deviation of the variance <sup>11</sup>; see Fig. 3.2.

Figure 3.2: Normalised mean (left panel) and variance (right panel) of each parameter in our model, as functions of iterations. The normalisation values are the mean and variance of the ensemble of particles positions at the last iteration. Red lines show one and two sigma levels of these normalisation values.

Once all parameters have entered the equilibrium state, we stop *emcee* by using the criterium of Gong & Flegal (2016) <sup>12</sup>. We chose this criterium because it was developed for high-dimensional problems and tested on Hierarchical Bayesian Models. In this criterium, the MCMC chain stops once its "effective sample size" (ESS, the size that an independent and identically distributed sample must have to provide the same inference) is larger than a minimum sample size computed using the required accuracy,  $v$ , for each parameter confidence interval  $(1 - \delta)100\%$ . Our *emcee* run stops once the ESS of the ensemble of walkers is greater than the minimum sample size needed for the required accuracy  $v = 0.05$  on the 68% confidence interval ( $\delta = 0.32$ ) of each parameter.

Description of the techniques used to obtain samples from the posterior distribution.

History and used versions.

---

<sup>11</sup>  $sd(\sigma^2) = \sigma^2 \sqrt{\kappa/n + 2/(n-1)}$  with  $\kappa$  the kurtosis and  $n$  the sample size.

<sup>12</sup> Implemented in the R package *mcmcse* (Flegal et al. 2016)

### **3.5.1 PSO**

#### **3.5.1.1 The charged PSO**

### **3.5.2 MCMC**

#### **3.5.2.1 Generalities**

#### **3.5.2.2 Flavours**

HMC,NUTS,Gibbs, Metropolis-Hasting, Affine invariant, stretch-move, Multi-Nest. It must be clear why we choose emcee and multinest

#### **3.5.2.3 Convergence**

#### **3.5.2.4 The evidence**

## **3.6 Codes**

#### **3.6.0.5 The modified charged PSO**

#### **3.6.0.6 Improvements of emcee**

#### **3.6.0.7 The GMM with missing values**

### **3.6.1 Parallel implementations**

Description of the implementation. MPI, python stan, etc. explain in detail the difficulties faced at implementing the different codes in the different servers.

# Chapter 4

## Results

### 4.1 Performance of the classifier

Compare and explain the differences with Stauffer members.

### 4.2 Velocity distribution

### 4.3 Spatial distribution

### 4.4 Luminosity distribution

### 4.5 Mass distribution

#### 4.5.1 The mass-luminosity relation

### 4.6 The mass distribution on time

### 4.7 The phase space

### 4.8 Updating the previous knowledge

## Chapter 5

### Conclusions and Future Work

# Appendix A

## Appendix Chapter Title

In the following we include ....





- Blackwell, T. & Bentley, P. 2002, in *Proceedings of the 4th Annual Conference on Genetic and Evolutionary Computations*, ed. W. Langdon, E. Cantu-Paz, K. Mathias, R. Roy, & D. Davis (San Francisco, CA, USA: Morgan kaufmann Publishers Inc.), 19–26
- Bouvier, J., Kendall, T., Meeus, G., et al. 2008, *A&A*, 481, 661
- Bouy, H., Bertin, E., Barrado, D., et al. 2015a, *A&A*, 575, A120
- Bouy, H., Bertin, E., Moraux, E., et al. 2013, *A&A*, 554, A101
- Bouy, H., Bertin, E., Sarro, L. M., et al. 2015b, *A&A*, 577, A148
- Bovy, J., Hogg, D. W., & Roweis, S. T. 2011, *The Annals of Applied Statistics*, 5, 1657
- Buchner, J., Georgakakis, A., Nandra, K., et al. 2014, *A&A*, 564, A125
- Cardelli, J. A., Clayton, G. C., & Mathis, J. S. 1989, *ApJ*, 345, 245
- Chabrier, G. 2003, *PASP*, 115, 763
- Chabrier, G. 2005, in *Astrophysics and Space Science Library*, Vol. 327, *The Initial Mass Function 50 Years Later*, ed. E. Corbelli, F. Palla, & H. Zinnecker, 41
- Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., & Dorie, V. 2015a, *Journal of Educational and Behavioral Statistics*, 40, 136
- Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., & Dorie, V. 2015b, *Journal of Educational and Behavioral Statistics*, 40, 136
- Clerc, M. & Kennedy, J. 2002, *Trans. Evol. Comp*, 6, 58
- Converse, J. M. & Stahler, S. W. 2008, *ApJ*, 678, 431
- D’Antona, F. 1998, in *Astronomical Society of the Pacific Conference Series*, Vol. 142, *The Stellar Initial Mass Function (38th Herstmonceux Conference)*, ed. G. Gilmore & D. Howell, 157
- Demory, B.-O. 2014, *ApJ*, 789, L20

- Diard, J. & Bessiere, P. 2008, Probabilistic reasoning and decision making in sensory-motor systems, 153
- Dirac, P. A. M. 1942, Proceedings of the Royal Society of London Series A, 180, 1
- Duquennoy, A. & Mayor, M. 1991, A&A, 248, 485
- Faherty, J. K., Burgasser, A. J., Walter, F. M., et al. 2012, ApJ, 752, 56
- Feeney, S. M., Johnson, M. C., McEwen, J. D., Mortlock, D. J., & Peiris, H. V. 2013, Phys. Rev. D, 88, 043012
- Fei-Fei, L. & Perona, P. 2005, in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 2, Ieee, 524–531
- Flegal, J. M., Hughes, J., & Vats, D. 2016, mcmcse: Monte Carlo Standard Errors for MCMC, Riverside, CA and Minneapolis, MN, r package version 1.2-1
- for Guides in Metrology, J. C. 2008, JCGM 100: Evaluation of Measurement Data - Guide to the Expression of Uncertainty in Measurement, Tech. rep., JCGM
- Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, Publications of the Astronomical Society of the Pacific, 125, 306
- Fraley, C. & Raftery, A. E. 2002, Journal of the American Statistical Association, 97, 611
- Fraley, C., Raftery, A. E., Murphy, T. B., & Scrucca, L. 2012, mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation
- Franklin, J. 2002, The Science of Conjecture: Evidence and Probability Before Pascal, A Johns Hopkins paperback (Johns Hopkins University Press)
- Gagné, J., Lafrenière, D., Doyon, R., Malo, L., & Artigau, É. 2014, ApJ, 783, 121
- Galli, P. A. B., Moraux, E., Bouy, H., et al. 2017, A&A, 598, A48
- Gelman, A. 2006, Bayesian Analysis, 1, 515

- Gelman, A. 2012, CHANCE, 25, 52
- Gelman, A., Carlin, J., Stern, H., et al. 2013, Bayesian Data Analysis, third edition edn. (CRC)
- Gelman, A. & Hill, J. 2007, Data Analysis Using Regression and Multilevel/Hierarchical Models, Analytical Methods for Social Research (Cambridge University Press)
- Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. 2008, Ann. Appl. Stat., 2, 1360
- Gong, L. & Flegal, J. M. 2016, Journal of Computational and Graphical Statistics, 25, 684
- González-Farias, G., Dominguez-Molina, A., & Gupta, A. K. 2004, Journal of Statistical Planning and Inference, 126, 521
- Good, I. J. 1980, Trabajos de Estadística Y de Investigación Operativa, 31, 489
- Goodman, J. & Weare, J. 2010, Communications in Applied Mathematics and Computational Science, 5, 65
- Gupta, A. K., González-Farias, G., & Dominguez-Molina, J. 2004, Journal of Multivariate Analysis, 89, 181
- Guthrie, B. N. G. 1987, QJRAS, 28, 289
- Hand, D. J. & Yu, K. 2001, International Statistical Review / Revue Internationale de Statistique, 69, 385
- Hogg, D. W., Bovy, J., & Lang, D. 2010a, ArXiv e-prints
- Hogg, D. W., Myers, A. D., & Bovy, J. 2010b, ApJ, 725, 2166
- Hong, D., Balzano, L., & Fessler, J. A. 2016, Towards a theoretical analysis of PCA for heteroscedastic data, arxiv 1610.03595
- Huang, A. & Wand, M. P. 2013, Bayesian Analysis, 8, 439

- Jaynes, E. 2003, *Probability Theory: The Logic of Science* (Cambridge University Press)
- Jefferys, T. R., Jefferys, W. H., Barnes, III, T. G., & Dambis, A. 2007, in *Astronomical Society of the Pacific Conference Series*, Vol. 371, *Statistical Challenges in Modern Astronomy IV*, ed. G. J. Babu & E. D. Feigelson, 433
- Kass, R. E. & Raftery, A. E. 1995, *Journal of the American Statistical Association*, 90, 773
- Koller, D. & Friedman, N. 2009, *Probabilistic Graphical Models: Principles and Techniques*, *Adaptive computation and machine learning* (MIT Press)
- Krone-Martins, A. & Moitinho, A. 2014, *A&A*, 561, A57
- Lindstrom, M. 1999, *Journal of Computational and Graphical Statistics*, 8, 333
- Liu, Z., Chen, Y., Tian, S., & Xu, Z. 2015, *Journal of Information and Computational Science*, 12, 775
- Lodieu, N., Deacon, N. R., & Hambly, N. C. 2012, *MNRAS*, 422, 1495
- Malo, L., Doyon, R., Lafrenière, D., et al. 2013, *ApJ*, 762, 88
- March, M. C., Karpenka, N. V., Feroz, F., & Hobson, M. P. 2014, *MNRAS*, 437, 3298
- McMichael, D. W. 1996, in *Proc. Fourth International Symposium on Signal Processing and its Applications (ISSPA)*, 377–378
- Morau, E., Bouvier, J., & Stauffer, J. R. 2001, *A&A*, 367, 211
- Morris, C. N. 1983, *Journal of the American Statistical Association*, 78, 47
- Muench, A. A., Lada, E. A., Lada, C. J., & Alves, J. 2002, *ApJ*, 573, 366
- Ozerov, A., Lagrange, M., & Vincent, E. 2013, *Computer Speech and Language*, 27, 874, special Issue on Speech Separation and Recognition in Multisource Environments

- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, *Journal of Machine Learning Research*, 12, 2825
- Press, W. H. 1997, in *Unsolved Problems in Astrophysics*, ed. J. N. Bahcall & J. P. Ostriker, 49–60
- Pugachev, V. 1965, *Theory of Random Functions and its Application to Control Problems* (Pergamon), 852
- Riedel, A. R., Blunt, S. C., Lambrides, E. L., et al. 2017, *AJ*, 153, 95
- Roberts, G. O. & Rosenthal, J. S. 2004, *Probability Surveys*, 1, 20
- Sale, S. E. 2012, *MNRAS*, 427, 2119
- Sampedro, L. & Alfaro, E. J. 2016, *MNRAS*, 457, 3949
- Sarro, L. M., Bouy, H., Berihuete, A., et al. 2014, *Astronomy & Astrophysics*, 14
- Schwarz, G. 1978, *Ann. Statist.*, 6, 461
- Shariff, H., Dhawan, S., Jiao, X., et al. 2016, *MNRAS*, 463, 4311
- Shkedy, Z., Decin, L., Molenberghs, G., & Aerts, C. 2007, *MNRAS*, 377, 120
- Smyth, G. K. 1998, in *Encyclopedia of Biostatistics*, ed. Armitage, P. & Colton, T. (John Wiley and Sons)
- Spiriti, S., Eubank, R., Smith, P. W., & Young, D. 2013, *Journal of Statistical Computation and Simulation*, 83, 1020
- Stauffer, J. R., Hartmann, L. W., Fazio, G. G., et al. 2007, *ApJS*, 172, 663
- Stauffer, J. R., Schultz, G., & Kirkpatrick, J. D. 1998, *ApJ*, 499, L199
- Stoneking, C. J. 2014
- Takeda, Y., Hashimoto, O., & Honda, S. 2017, *PASJ*, 69, 1
- Tapiador, D., Berihuete, A., Sarro, L. M., Julbe, F., & Huedo, E. 2017, *Astronomy and Computing*, 19, 1

Thies, I. & Kroupa, P. 2007, ApJ, 671, 767

Trumpler, R. J. 1921, Lick Observatory Bulletin, 10, 110

Wolfgang, A. & Lopez, E. 2015, ApJ, 806, 183